EXHIBIT B

Combining computational and experimental screening for rapid optimization of protein properties

Robert J. Hayes*, Jörg Bentzien*, Marie L. Ary, Marian Y. Hwang, Jonathan M. Jacinto, Jost Vielmetter, Anirban Kundu, and Bassil I. Dahiyat†

Xencor, 111 West Lemon Avenue, Monrovia, CA 91016

Communicated by Pamela J. Bjorkman, California Institute of Technology, Pasadena, CA, October 16, 2002 (received for review August 14, 2002)

We present a combined computational and experimental method for the rapid optimization of proteins. Using β-lactamase as a test case, we redesigned the active site region using our Protein Design Automation technology as a computational screen to search the entire sequence space. By eliminating sequences incompatible with the protein fold, Protein Design Automation rapidly reduced the number of sequences to a size amenable to experimental screening, resulting in a library of ≈200,000 mutants. These were then constructed and experimentally screened to select for variants with improved resistance to the antibiotic cefotaxime. In a single round, we obtained variants exhibiting a 1,280-fold increase in resistance. To our knowledge, all of the mutations were novel, i.e., they have not been identified as beneficial by random mutagenesis or DNA shuffling or seen in any of the naturally occurring TEM β-lactamases, the most prevalent type of Gram-negative β-lactamases. This combined approach allows for the rapid improvement of any property that can be screened experimentally and provides a powerful broadly applicable tool for protein engineering.

computational protein design | protein engineering | mutagenesis | directed evolution | β-lactamase

The increased use of enzymes and other proteins in the chemical, agricultural, and pharmaceutical industries has generated considerable interest in the design of proteins with new and improved properties. Two different but complementary technologies have been applied to this goal: (*i*) rational design, which relies on structural and mechanistic knowledge and human expertise; and (*ii*) directed evolution methods such as error-prone PCR, phage display, and DNA shuffling, which use random mutagenesis or recombination to create diversity and then experimentally screen the libraries generated for desired properties (1). Directed evolution has been successfully used on a wide range of proteins (2–7). However, this approach is limited by the number of sequences that can be screened experimentally (about 10^{14} for library panning and 10^7 for high-throughput screening). Rational design has also been applied with some success (8–10), but it was not until computational methods were developed that it could be used comprehensively.

Computational techniques use protein design algorithms to perform *in silico* screening of protein sequences (11–17). By taking advantage of the speed of computers, these methods allow a vast number of sequences to be screened ($\approx 10^{80}$). The ability to search such large sequence spaces drastically increases the possibility of finding novel proteins with improved properties. Computational techniques have also been developed to enhance the efficiency of directed evolution methods (18, 19).

One computational design tool that has proven effective is Protein Design Automation (PDA) (13). PDA begins with the three-dimensional structural model of the protein to be designed and predicts the optimal sequence that will adopt this fold, allowing all or a specified set of residues to change. The fitness of sequences is scored by using physical potential functions that model the energetic interactions of protein atoms (20); stable low-energy sequences are given the best scores. By using extremely efficient search algorithms, up to 10^{80} sequences can be accurately screened within hours (21–23). Multiple simultaneous mutations can be made, and novel sequences that are very different from wild type can be discovered. PDA has shown tremendous success in designing proteins with improved stability and conformational specificity (13, 14, 24–28) and has even been used to engineer a catalytic site into a previously nonreactive protein (29).

In these studies, only a few optimal sequences calculated by PDA were made and tested experimentally. The utility of PDA can be extended significantly, however, if it is used to generate a library of sequences, all of which are predicted to be stable and fold into a predetermined structure. Unlike random libraries, where most of the mutations are deleterious, the mutant sequences in the PDA library are computationally screened to eliminate destabilizing mutations and sequences inconsistent with the proper fold. The selected sequences are then experimentally screened for desired properties such as improved catalytic activity, substrate specificity, or receptor binding. Therefore, PDA is a computational prescreen to decrease the sequence space many orders of magnitude, while maintaining broad diversity, to a number easily amenable to experimental screening. By coupling PDA with experimental screening, we combine the advantages of computational design with those of directed evolution: namely, access to a vast sequence space and the ability to improve any protein property that can be captured by a screen.

In this paper, we demonstrate the feasibility of this approach by using it to increase the resistance of bacteria toward the antibiotic cefotaxime by optimizing TEM-1 β-lactamase, the most prevalent plasma-encoded β-lactamase in Gram-negative bacteria.

Methods

Structure Preparation. The crystal structure of TEM-1 β-lactamase (Protein Data Bank no. 1BTL) (30) was used as the starting point for modeling. All water molecules and the sulfate group were removed; the side chains of residues N132, N154, N170, H122, and H289 were flipped to form a better hydrogen bond network; and the disulfide bond between C77 and C123 was formed manually. The program BIOGRAF (Molecular Simulations, San Diego) was used to generate explicit hydrogens, and 50 steps of conjugate gradient minimization were performed by using the Dreiding II force field (31) without the electrostatics term. The minimization is done to make the structure compatible with our force-field parameters and results in very slight changes to the coordinates.

Construction of Mutant Library. To facilitate introduction of the mutations into the *TEM-1* gene, a pCR-Blunt (Invitrogen) vector containing the *TEM-1* gene was digested with *Xba*I and *Hin*dIII,

Abbreviations: PDA, Protein Design Automation; MIC, minimum inhibitory concentration; GMEC, global minimum energy conformation.

*R.J.H. and J.B. contributed equally to this work.

†To whom correspondence should be addressed. E-mail: baz@xencor.com.

www.pnas.org/cgi/doi/10.1073/pnas.212627499

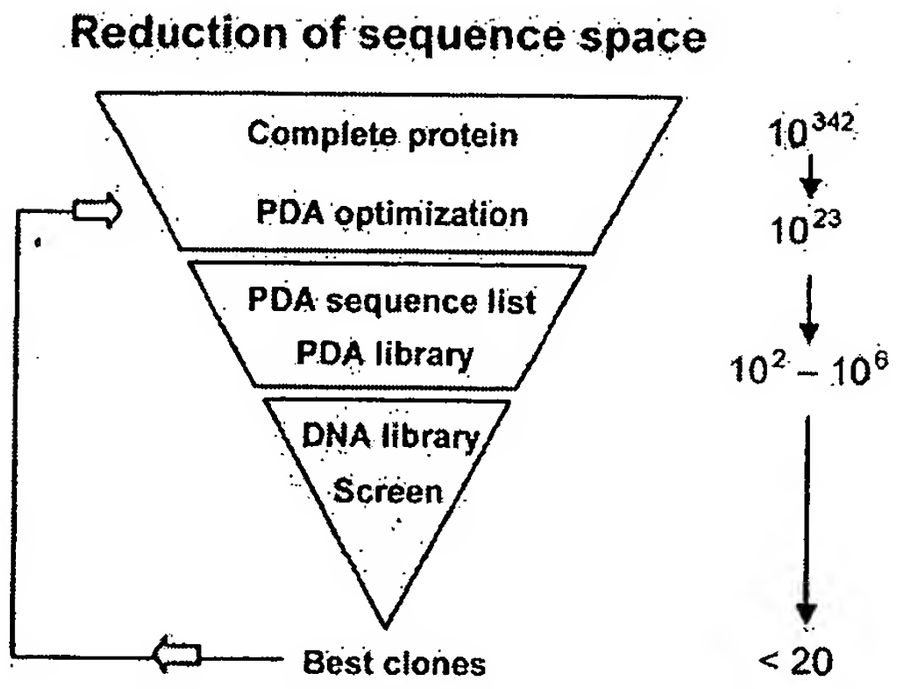

Fig. 2. Reduction of sequence space for PDA design of TEM-1 β-lactamase. Computational screening with PDA and judicious application of cutoffs reduced the sequence space 18 orders of magnitude for the 19 residues explicitly considered and more than 300 orders of magnitude for the entire protein. This conformational screen specified a library for experimental screening of ≈200,000 mutant sequences, enriched for structural integrity.

amino acid occurrences at each of the designed positions. Different cutoffs or weighting functions can be applied to define a library of a desired size, appropriate for experimental screening. Structure or sequence alignment information, experimental data, and diversity considerations may also be taken into account in defining the mutant library.

Recursive PCR with overlapping oligonucleotides is then used to synthesize the genes containing all of the mutant sequences in the PDA-defined library. The genes are pooled and cloned, and the mutant proteins are expressed in an appropriate host such as *E. coli*. The mutant proteins are screened experimentally for desired properties, and the best mutants are isolated and characterized. These results can be used as feedback for additional rounds of computational design, library generation, and screening.

Reduction of Sequence Space. The use of PDA as a computational screen allows us to access a vast sequence space and, by eliminating sequences predicted to be destabilizing or inconsistent with the proper fold, reduce it to a size amenable to experimental screening. The reduction of sequence space obtained for TEM-1 β-lactamase, our test case, is shown in Fig. 2. If we were to consider the entire β-lactamase protein (263 residues) and allow all 20 amino acids at each position, we would need to screen 20^{263} or $\approx 1.4 \times 10^{342}$ sequences. By focusing the design to a particular region (19 residues near the active site) and using a slightly restricted set of amino acids (19), we reduced this to 7×10^{23} sequences, a number that can easily be screened computationally, but not experimentally. We then chose cutoffs for the Monte Carlo list and the probability table that would define a library within the limits of experimental screening. In this case, we specified a library of ≈200,000 mutant sequences, a reduction of 18 orders of magnitude for the residues explicitly considered and an overall reduction of more than 300 orders of magnitude for the entire protein.

Computational Design of β-Lactamase. The hydrolysis of β-lactam antibiotics, catalyzed by β-lactamase, is a common mechanism by which bacteria become resistant to antibiotics (37). The most prevalent plasmid-encoded β-lactamase in Gram-negative bacteria is the class A TEM-1 β-lactamase (38). This enzyme hydrolyzes ampicillin efficiently but is inefficient at hydrolyzing the cephalosporin cefotaxime. Our goal was to use PDA to design β-lactamase variants that confer increased resistance toward cefotaxime.

Optimizing the area around the active site is likely to have a significant effect on enzyme activity and substrate specificity (37, 39). Although more distant mutations can also be effective (40), the rationale for how to select such positions is less obvious. Designing residues around the active site also serves as a stringent test of the ability of PDA to predict nondisruptive mutations. We therefore focused our design on residues within 5 Å of the active site residues S70, K73, S130, E166, and K234. These criteria resulted in 19 positions that were allowed to change: M69, T71, F72, V74, V103, Y105, A126, I127, N132, A135, N136, L169, N170, M211, D214, K234, S235, G236, and I247. All 20 amino acids, except cysteine and proline, were considered at these positions. The catalytic residues (S70, K73, S130, and E166) were not allowed to change their amino acid identities; however, their conformations could vary. An expanded version of the backbone-dependent rotamer library of Dunbrack and Karplus (41) was used in all of the calculations, and the DEE algorithm was used to find the GMEC. The computational details, residue classification, and potential functions used are described in previous work (13, 14, 20, 42).

Definition of Mutant Library. Optimization with PDA predicted an optimal sequence with nine mutations. Starting from this GMEC, we applied Monte Carlo simulated annealing to produce a rank-ordered list of the 1,000 lowest energy sequences. A probability table was generated from this list by counting the amino acid occurrences at each of the 19 designed positions (Table 1). A 10% cutoff was then applied to the probability table to define a library of mutant sequences for experimental screening; that is, for a given position, an amino acid identity was included in the library if it had a 10% or greater probability of occurrence. To ensure that the library spanned the complete sequence space from the wild-type enzyme to the most distantly related PDA mutant, we always included the wild-type identity at all designed positions, even if it did not appear in the Monte Carlo list. With a 10% cutoff, this gave us a library of 172,800 unique sequences; a 20% cutoff would have resulted in a much smaller library of 4,806.

Construction of Genes for Mutant Library. Recursive PCR with overlapping oligonucleotides was used to synthesize the *TEM-1* β-lactamase genes containing all 172,800 mutant sequences in the PDA library. Synthetic oligonucleotides containing the designed mutations were pooled to create desired diversity at each site. Two separate reactions were performed: one that contained only a proofreading DNA polymerase (*Pfu* DNA polymerase), termed the nonerror prone reaction, and one that contained both *Pfu* DNA polymerase and *Taq*DNA polymerase, termed the error-prone reaction. The mutated genes were cloned and transformed into *E. coli*.

Validation of Mutant Library. Sixty individual clones from the nonerror-prone library were sequenced by standard techniques. The plasmids contained intact ORFs with the desired mutations. No additional mutations were detected. With a sample size of 60, we were able to find all of the specified mutations at each designed position. It is impossible to find all combinations of the mutations within this small sample (the library contained 172,800 unique sequences), but none of the clones were identical and we were unable to detect a statistically significant bias toward any particular mutation at any position. This result indicates that we have developed an efficient method for converting a PDA-defined library into an experimental library containing all of the mutated genes required to encode the desired mutant sequences.

Experimental Screen for β-Lactamase Activity. Experimental libraries of ≈500,000 individual *E. coli* colonies expressing the mu-

in a single round, we were able to use very stringent selection conditions and directly obtain highly resistant variants. The identification of incrementally improved sequences was not necessary.

Substrate Specificity. We also measured resistance to ampicillin and found no growth at 100 μg/ml, significantly less than the MIC of 4,096 μg/ml reported for the wild type (43). This result suggests that our screens identified clones whose resistance to cefotaxime had dramatically improved, whereas their resistance to ampicillin was reduced at least 40-fold. The relative substrate specificity toward cefotaxime vs. ampicillin was thus enhanced 25,000- to 50,000-fold.

PDA Mutants Are Novel. The most active mutant from each library was isolated and sequenced. PDA-1 had eight mutations (M69D, V103Q, Y105N, N132M, L169A, N170L, S235D, and G236S), all designated in the PDA library (see Table 1). PDA-2 had five PDA-designed mutations (V103Q, Y105N, I127L, L169A, and G236S) and one random mutation (S235Y). The S235Y mutation was not predicted by PDA due to steric clashes. Protein backbone motion, which is required to relieve the clash, is not considered in the computation. None of the mutations in PDA-1 or -2 have been identified by full gene random mutagenesis or DNA shuffling studies (2, 39, 45, 46) or have been observed in the 105 naturally occurring TEM β-lactamases (G. Jacoby and K. Bush, www.lahey.org/studies/temtable.htm). Orencia *et al.* (47) discussed the emergence of antibiotic resistances in β-lactamases and showed that there is an overlap between the mutations discovered by directed evolution and those occurring in natural evolution. PDA, however, accesses the entire designed sequence space including all possible combinations of mutations and therefore can produce multiple simultaneous mutations. PDA is therefore more likely to identify novel mutants with desired properties. The lone random mutation in PDA-2 (S235Y) was in the active site region, suggesting that the novel context of the PDA-designed mutants allowed this previously unobserved, but beneficial, mutation to emerge.

Two of the mutations in PDA-1 (V105N and G236S) were reverted to wild type to create a backcross mutant (PDA-3). This PDA-3 sequence is present in the library defined with a 20% cutoff but is absent if a 10% cutoff is used (see Table 1). PDA-3 exhibited the same cefotaxime resistance as PDA-1 (Table 2), indicating that a smaller PDA-library (4,806 vs. 172,800 sequences) can also generate mutants with significantly improved activity. Additional backcrosses were done to examine the role of the other six mutations in PDA-1. No single mutation was primarily responsible for the improved resistance, and no simple additivity was apparent, suggesting that the mutations are coupled. This conclusion is supported by extensive replacement mutagenesis studies of three-residue segments around the active site (39). They found a mutant (E168G, L169A, and N170G) that included one of our mutations (L169A), but it showed only a marginal (2-fold) improvement in cefotaxime resistance. Although they also tested most of our other mutations, no increased resistance was found for any of these. This lack of improved resistance indicates that the broader context of many simultaneous mutations provided by our approach was required to find our highly active sequences.

Comparison with Other Mutants. To compare the activity of our PDA-designed mutants with those obtained in other studies, we introduced some previously reported mutations into our wild-type gene, including E104K/G238S (comparable to TEM-15) (2, 46) and A18V/E104K/M182T/G238S (comparable to ST-1) (2, 46). TEM-15 is a naturally occurring β-lactamase that is active against cefotaxime, and ST-1 is a highly active TEM-1 variant discovered from three rounds of DNA shuffling. We tested the ability of these mutants to confer resistance to cefotaxime. Wild type had a MIC of 0.1 μg/ml, comparable to the values reported by others; TEM-15 and ST-1 had MICs of 16 and 256 μg/ml, respectively, also in line with previously published work (Table 2) (2, 44, 46, 48–50).

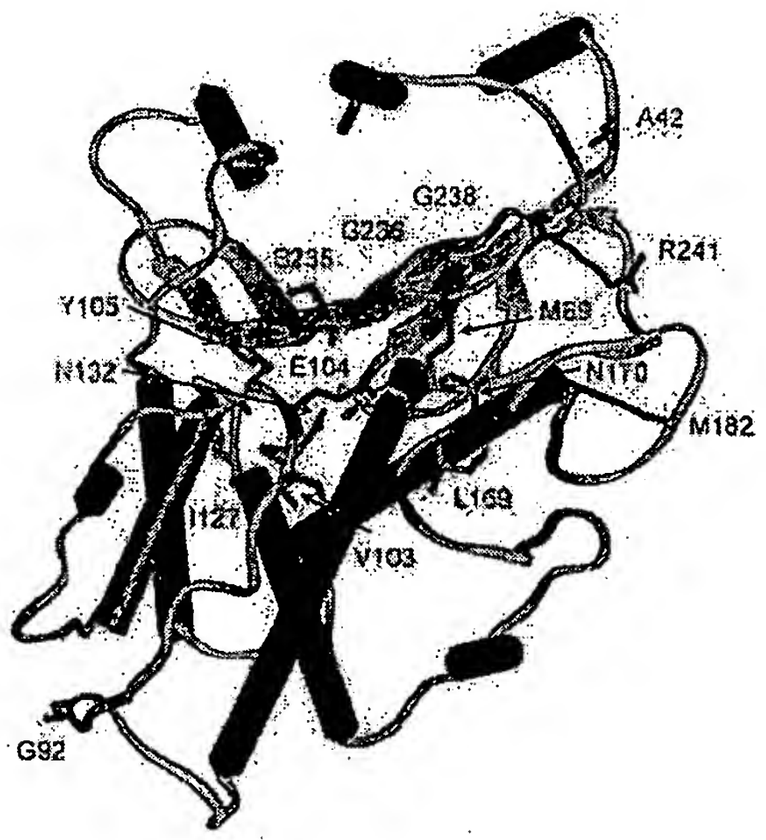

Fig. 3. Location of mutations in PDA-1 and -2 (green) vs. those obtained by DNA shuffling (2) and random hypermutagenesis (46) (magenta). The wild-type TEM-1 β-lactamase structure is illustrated, and the side chains of the mutated positions are shown. The catalytic serine (S70) is depicted in blue. The average distance between the C_α atoms and the catalytic nucleophile (O_γS70) in our PDA-1 and -2 mutations was 8.0 and 8.6 Å, respectively, vs. 16.0 Å for the mutations in ST-2 and -3 (Stemmer's best mutants) (2) and 12.1 Å for 3D.5 (Zaccolo and Gherardi's best mutant) (46). This difference in distances illustrates that the mutations found by PDA are near the designed active site area, whereas those found by DNA shuffling and random hypermutagenesis are farther away.

Location of Mutations. The mutations in all our variants are located in or near the active site, because our computational design restricted changes to this region. Directed evolution methods, however, tend to produce mutations spread over the entire protein structure. For example, almost all of the mutations in the best mutants obtained by DNA shuffling (2) and random hypermutagenesis (46) are located far from the active site (Fig. 3). It is possible that these techniques seldom produce mutations close to the active site, because they rely on incremental changes; a single change in the first round of screening must be beneficial to be passed to the second round. However, point mutations in the active site area are usually disruptive. Our approach, on the other hand, allows multiple simultaneous mutations in a single round, which can have compensating or even synergistic effects.

Sequence Space Coverage. Most of the mutations observed in our PDA variants require a minimum of two nucleotide changes, and one, M69D, can be made only by a triple nucleotide change (Table 3). Double- or triple-nucleotide changes within a single codon are very difficult to achieve by using random mutagenesis techniques such as error-prone PCR or single-gene DNA shuffling. This limitation is demonstrated by the fact that each of the mutations found in the directed evolution studies (2, 45, 46) as well as those observed in the 105 naturally occurring TEM variants (G. Jacoby and K. Bush, www.lahey.org/studies/temtable.htm) could be obtained by a single nucleotide change. If one considers all of the substitutions that are possible for each of the 20 amino acids, on average only seven can be achieved by a single nucleotide change. The sequence space coverage is further reduced by codon preferences, biases for transitions over transversions, and A $\leftrightarrow$ T over G $\leftrightarrow$ C mutations. These restrictions severely limit the sequence